

---

# Optimistic Agents are Asymptotically Optimal

---

Peter Sunehag and Marcus Hutter

peter.sunehag@anu.edu.au, marcus.hutter@anu.edu.au

Research School of Computer Science  
Australian National University  
Canberra, ACT, 0200, Australia

September 2012

## Abstract

We use optimism to introduce generic asymptotically optimal reinforcement learning agents. They achieve, with an arbitrary finite or compact class of environments, asymptotically optimal behavior. Furthermore, in the finite deterministic case we provide finite error bounds.

## Keywords

Reinforcement Learning; Optimism; Optimality; Agents; Uncertainty.

## 1 Introduction

This article studies a fundamental question in artificial intelligence; given a set of environments, how do we define an agent that eventually acts optimally regardless of which of the environments it is in. This question relates to the even more fundamental question of what intelligence is. [Hut05] defines an intelligent agent as one that can act well in a large range of environments. He studies arbitrary classes of environments with particular attention to universal classes of environments like all computable (deterministic) environments and all lower semi-computable (stochastic) environments. He defines the AIXI agent as a Bayesian reinforcement learning agent with a universal hypothesis class and a Solomonoff prior. This agent has some interesting optimality properties. Besides maximizing expected utility with respect to the a priori distribution by design, it is also Pareto optimal and self-optimizing when this is possible for the considered class. It was, however, shown in [Ors10] that it is not guaranteed to be asymptotically optimal for all computable (deterministic) environments. [LH11a] shows that this is not surprising since, at least for geometric discounting, no agent can be. [LH11a] also shows that in a weaker (in average) sense, optimality can be achieved for the class of all computable environments using

an algorithm that includes long exploration phases. Furthermore, it is simple to realize that Bayesian agents do not always achieve optimality for a finite class of deterministic environments even if all prior weights are strictly positive.

We use the principle of optimism to define an agent that for any finite class of deterministic environments, eventually acts optimally. We extend our results to the case of finite and compact classes of stochastic environments. In the deterministic case we also prove finite error bounds. Optimism has previously been used to design exploration strategies for both discounted and undiscounted MDPs [KS98, SL05, AO06, LH12], though here we define optimistic algorithms for any finite class of environments.

**Related work.** Besides AIXI [Hut05] that was discussed above, [LH11a] introduces an agent which achieves asymptotic optimality in an average sense for the class of all deterministic computable environments. There is, however, no time step after which it is optimal at every time step. This is due to an infinite number of long exploration phases. We introduce an agent, that for finite classes of environments, does eventually achieve optimality for every time step. For the stochastic case, the agent achieves with any given probability, optimality within  $\epsilon$  for any  $\epsilon > 0$ . Our very simple agent is relying elegantly on the principle of optimism, used previously in the restrictive MDP case with discounting [KS98, SL05, LH12] and without [AO06], instead of an indefinite number of explicitly enforced bursts of exploration. [RH08] also introduces an agent that relies on bursts of exploration with the aim of achieving asymptotic optimality. The asymptotic optimality guarantees are restricted to a setting where all environments satisfy a certain restrictive value-preservation property. [EDKM05] studied learning general Partially Observable Markov Decision Processes (POMDPs). Though POMDPs constitute a very general reinforcement learning setting, we are interested in agents that can be given any (deterministic or stochastic) class of environments and successfully utilize the knowledge that the true environment lies in this class.

**Background.** We will consider an agent [RN10, Hut05] that interacts with an environment through performing actions  $a_t$  from a finite set  $\mathcal{A}$  and receives observations  $o_t$  from a finite set  $\mathcal{O}$  and rewards  $r_t$  from a finite set  $\mathcal{R} \subset [0, 1]$ . Let  $\mathcal{H} = (\mathcal{A} \times \mathcal{O} \times \mathcal{R})^*$  be the set of histories and  $R : \mathcal{H} \rightarrow \mathbb{R}$  the return

$$R(a_1 o_1 r_1 a_2 o_2 r_2 \dots a_n o_n r_n) = \sum_{j=1}^n r_j \gamma^j$$

with the obvious extension to infinite sequences. A function from  $\mathcal{H} \times \mathcal{A}$  to  $\mathcal{O} \times \mathcal{R}$  is called a deterministic environment (studied in Section 2). A function  $\pi : \mathcal{H} \rightarrow \mathcal{A}$  is called a policy or an agent. We define the value function  $V$  by  $V_\nu^\pi(h_{t-1}) := R(h_{t:\infty}) = \sum_{i=t}^\infty \gamma^{i-t} r_i$  where the sequence  $r_i$  are the rewards achieved by following  $\pi$  from time step  $t$  onwards in environment  $\nu$  after having seen  $h_{t-1}$ .

Instead of viewing the environment as a function from  $\mathcal{H} \times \mathcal{A}$  to  $\mathcal{O} \times \mathcal{R}$  we can equivalently write it as a function  $\nu : \mathcal{H} \times \mathcal{A} \times \mathcal{O} \times \mathcal{R} \rightarrow \{0, 1\}$  where we

write  $\nu(o, r|h, a)$  for the function value of  $(h, a, o, r)$ . It equals zero if in the first formulation  $(h, a)$  is not sent to  $(o, r)$  and 1 if it is. In the case of stochastic environments, which we will study in Section 3, we instead have a function  $\nu : \mathcal{H} \times \mathcal{A} \times \mathcal{O} \times \mathcal{R} \rightarrow [0, 1]$  such that  $\sum_{o,r} \nu(o, r|h, a) = 1 \ \forall h, a$ . Furthermore, we define  $\nu(h_t|\pi) := \nu(or_{1:t}|\pi) := \prod_{i=1}^t \nu(o_i r_i|a_i, h_{i-1})$  where  $a_i = \pi(h_{i-1})$ .  $\nu(\cdot|\pi)$  is a probability measure over strings or sequences as will be discussed in the next section and we can define  $\nu(\cdot|\pi, h_{t-1})$  by conditioning  $\nu(\cdot|\pi)$  on  $h_{t-1}$ . We define  $V_\nu^\pi(h_{t-1}) := \mathbb{E}_{\nu(\cdot|\pi, h_{t-1})} R(h_{t:\infty})$  as the  $\nu$ -expected return of policy  $\pi$ .

A special case of an environment is a Markov Decision Process (MDP) [SB98]. This is the classical setting for reinforcement learning. In this case the environment does not depend on the full history but only on the latest observation and action and is, therefore, a function from  $\mathcal{O} \times \mathcal{A} \times \mathcal{O} \times \mathcal{R}$  to  $[0, 1]$ . In this situation one often refers to the observations as states since the latest observation tells us everything we need to know. In this situation, there is an optimal policy that can be represented as a function from the state set  $\mathcal{S} (:= \mathcal{O})$  to  $\mathcal{A}$ . We only need to base our decision on the latest observation. Several algorithms [KS98, SL05, LH12] have been devised for solving discounted ( $\gamma < 1$ ) MDPs for which one can prove PAC (Probably Approximately Correct) bounds. They are finite time bounds that hold with high probability and depend only polynomially on the number of states, actions and the discount factor. These methods are relying on optimism as the method for making the agent sufficiently explorative. Optimism roughly means that one has high expectations for what one does not yet know. Optimism was also used to prove regret bounds for undiscounted ( $\gamma = 1$ ) MDPs in [AO06] which was extended to feature MDPs in [MMR11]. Note that these methods are restricted to MDPs and that we do not make any (Markov, ergodicity, stationarity, etc.) assumptions on the environments, only on the size of the class.

**Outline.** In this article we will define optimistic agents in a far more general setting than MDPs and prove asymptotic optimality results. The question of their mere existence is already non-trivial, hence asymptotic results deserve attention. In Section 2 we consider finite classes of deterministic environments and introduce a simple optimistic agent that is guaranteed to eventually act optimally. We also provide finite error bounds. In Section 3 we generalize to finite classes of stochastic environments and in Section 4 to compact classes.

## 2 Finite Classes of Deterministic Environments

Given a finite class of deterministic environments  $\mathcal{M} = \{\nu_1, \dots, \nu_m\}$ , we define an algorithm that for any unknown environment from  $\mathcal{M}$  eventually achieves optimal behavior in the sense that there exists  $T$  such that maximum reward is achieved from time  $T$  onwards. The algorithm chooses an optimistic hypothesis from  $\mathcal{M}$  in the sense that it picks the environment in which one can achieve the highest reward (in case of a tie, choose the environment which comes first in an enumeration of  $\mathcal{M}$ ) and then the policy that is optimal for this environment is

followed. If this hypothesis is contradicted by the feedback from the environment, a new optimistic hypothesis is picked from the environments that are still consistent with  $h$ . This technique has the important consequence that if the hypothesis is not contradicted we are still acting optimally when optimizing for this incorrect hypothesis.

**Require:** Finite class of deterministic environments  $\mathcal{M}_0 \equiv \mathcal{M}$

```

1:  $t = 1$ 
2: repeat
3:    $(\pi^*, \nu^*) \in \arg \max_{\pi \in \Pi, \nu \in \mathcal{M}_{t-1}} V_\nu^\pi(h_{t-1})$ 
4:   repeat
5:      $a_t = \pi^*(h_{t-1})$ 
6:     Perceive  $o_t r_t$  from environment  $\mu$ 
7:      $h_t \leftarrow h_{t-1} a_t o_t r_t$ 
8:     Remove all inconsistent environments from  $\mathcal{M}_t$ 
       ( $\mathcal{M}_t := \{\nu \in \mathcal{M}_{t-1} : h_t^{\pi^\circ, \nu} = h_t\}$ )
9:      $t \leftarrow t + 1$ 
10:  until  $\nu^* \notin \mathcal{M}_{t-1}$ 
11: until  $\mathcal{M}$  is empty

```

**Algorithm 1:** Optimistic Agent ( $\pi^\circ$ ) for Deterministic Environments

Let  $h_t^{\pi, \nu}$  be the history up to time  $t$  generated by policy  $\pi$  in environment  $\nu$ . In particular let  $h^\circ := h^{\pi^\circ, \mu}$  be the history generated by Algorithm 1 (policy  $\pi^\circ$ ) interacting with the actual “true” environment  $\mu$ . At the end of cycle  $t$  we know  $h_t^\circ = h_t$ . An environment  $\nu$  is called consistent with  $h_t$  if  $h_t^{\pi^\circ, \nu} = h_t$ . Let  $\mathcal{M}_t$  be the environments consistent with  $h_t$ . The algorithm only needs to check whether  $o_t^{\pi^\circ, \nu} = o_t$  and  $r_t^{\pi^\circ, \nu} = r_t$  for each  $\nu \in \mathcal{M}_{t-1}$ , since previous cycles ensure  $h_{t-1}^{\pi^\circ, \nu} = h_{t-1}$  and trivially  $a_t^{\pi^\circ, \nu} = a_t$ . The maximization in Algorithm 1 that defines optimism at time  $t$  is performed over all  $\nu \in \mathcal{M}_t$ , the set of consistent hypotheses at time  $t$ , and  $\pi \in \Pi = \Pi^{all}$  is the class of *all* deterministic policies.

**Theorem 1** (Optimality, Finite Deterministic Class). *If we use Algorithm 1 ( $\pi^\circ$ ) in an environment  $\mu \in \mathcal{M}$ , then there is  $T < \infty$  such that*

$$V_\mu^{\pi^\circ}(h_t) = \max_{\pi} V_\mu^\pi(h_t) \quad \forall t \geq T.$$

A key to proving Theorem 1 is time-consistency [LH11b] of geometric discounting. The following lemma tells us that if we act optimally with respect to a chosen optimistic hypothesis, it remains optimistic until contradicted.

**Lemma 2** (Time-consistency). *Suppose  $(\pi^*, \nu^*) \in \arg \max_{\pi \in \Pi, \nu \in \mathcal{M}_t} V_\nu^\pi(h_t)$ , that we act according to  $\pi^*$  from time  $t$  to time  $\tilde{t}-1$  and that  $\nu^*$  is still consistent at time  $\tilde{t} > t$ , then  $(\pi^*, \nu^*) \in \arg \max_{\pi \in \Pi, \nu \in \mathcal{M}_{\tilde{t}}} V_\nu^\pi(h_{\tilde{t}})$ .*

*Proof.* Suppose that  $V_{\nu^*}^{\pi^*}(h_{\tilde{t}}) < V_{\tilde{\nu}}^{\tilde{\pi}}(h_{\tilde{t}})$  for some  $\tilde{\pi}, \tilde{\nu}$ . It holds that  $V_{\nu^*}^{\pi^*}(h_t) = C + \gamma^{\tilde{t}-t} V_{\nu^*}^{\pi^*}(h_{\tilde{t}})$  where  $C$  is the accumulated reward between  $t$  and  $\tilde{t}-1$ . Let

$\hat{\pi}$  be a policy that equals  $\pi^*$  from  $t$  to  $\tilde{t} - 1$  and then equals  $\tilde{\pi}$ . It follows that  $V_{\tilde{\nu}}^{\hat{\pi}}(h_t) = C + \gamma^{t-t} V_{\tilde{\nu}}^{\hat{\pi}}(h_{\tilde{t}}) > C + \gamma^{t-t} V_{\nu^*}^{\pi^*}(h_{\tilde{t}}) = V_{\nu^*}^{\pi^*}(h_t)$  which contradicts the assumption  $(\pi^*, \nu^*) \in \arg \max_{\pi \in \Pi, \nu \in \mathcal{M}_t} V_{\nu}^{\pi}(h_t)$ . Therefore,  $V_{\nu^*}^{\pi^*}(h_{\tilde{t}}) \geq V_{\tilde{\nu}}^{\tilde{\pi}}(h_{\tilde{t}})$  for all  $\tilde{\pi}, \tilde{\nu}$ . ■

*Proof. (Theorem 1)* At time  $t$  we know  $h_t$ . If some  $\nu \in \mathcal{M}_{t-1}$  is inconsistent with  $h_t$ , i.e.  $h_t^{\pi^\circ, \nu} \neq h_t$ , it gets removed, i.e. is not in  $\mathcal{M}_{t'}$  for all  $t' \geq t$ .

Since  $\mathcal{M}_0 = \mathcal{M}$  is finite, such inconsistencies can only happen finitely often, i.e. from some  $T$  onwards we have  $\mathcal{M}_t = \mathcal{M}_\infty$  for all  $t \geq T$ . Since  $h_t^{\pi^\circ, \mu} = h_t \forall t$ , we know that  $\mu \in \mathcal{M}_t \forall t$ .

Assume  $t \geq T$  henceforth. The optimistic hypothesis will not change after this point. If the optimistic hypothesis is the true environment  $\mu$ , we have obviously chosen the true optimal policy.

In general, the optimistic hypothesis  $\nu^*$  is such that it will never be contradicted while actions are taken according to  $\pi^\circ$ , hence  $(\pi^*, \nu^*)$  do not change anymore. This implies

$$V_{\mu}^{\pi^\circ}(h_t) = V_{\mu}^{\pi^*}(h_t) = V_{\nu^*}^{\pi^*}(h_t) = \max_{\nu \in \mathcal{M}_t} \max_{\pi \in \Pi} V_{\nu}^{\pi}(h_t) \geq \max_{\pi \in \Pi} V_{\mu}^{\pi}(h_t)$$

for all  $t \geq T$ . The first equality follows from  $\pi^\circ$  equals  $\pi^*$  from  $t \geq T$  onwards. The second equality follows from consistency of  $\nu^*$  with  $h_{1:\infty}^\circ$ . The third equality follows from optimism, the constancy of  $\pi^*, \nu^*$ , and  $\mathcal{M}_t$  for  $t \geq T$ , and time-consistency of geometric discounting (Lemma 2). The last inequality follows from  $\mu \in \mathcal{M}_t$ . The reverse inequality  $V_{\mu}^{\pi^*}(h_t) \leq \max_{\pi} V_{\mu}^{\pi}(h_t)$  follows from  $\pi^* \in \Pi$ . Therefore  $\pi^\circ$  is acting optimally at all times  $t \geq T$ . ■

Besides the eventual optimality guarantee above, we also provide a bound on the number of time steps for which the value of following Algorithm 1 is more than a certain  $\varepsilon > 0$  less than optimal. The reason this bound is true is that we only have such suboptimality for a certain number of time steps before a point where the current hypothesis becomes inconsistent and the number of such inconsistency points are bounded by the number of environments.

**Theorem 3** (Finite error bound). *Following  $\pi^\circ$  (Algorithm 1),*

$$V_{\mu}^{\pi^\circ}(h_t) \geq \max_{\pi \in \Pi} V_{\mu}^{\pi}(h_t) - \varepsilon, \quad 0 < \varepsilon < 1/(1 - \gamma)$$

*for all but at most  $|\mathcal{M}|^{\frac{\log \varepsilon (1-\gamma)}{\gamma-1}}$  time steps  $t$ .*

*Proof.* Consider the  $\ell$ -truncated value

$$V_{\nu, \ell}^{\pi}(h_t) := \sum_{i=t+1}^{t+\ell} \gamma^{i-t-1} r_i$$

where the sequence  $r_i$  are the rewards achieved by following  $\pi$  from time  $t+1$  to  $t+\ell$  in  $\nu$  after seeing  $h_t$ . By letting  $\ell = \frac{\log \varepsilon(1-\gamma)}{\log \gamma}$  (which is positive due to negativity of both numerator and denominator) we achieve  $|V_{\nu,\ell}^\pi(h_t) - V_\nu^\pi(h_t)| \leq \frac{\gamma^\ell}{1-\gamma} = \varepsilon$ . Let  $(\pi_t^*, \nu_t^*)$  be the policy-environment pair selected by Algorithm 2 in cycle  $t$ .

Let us first assume  $h_{t+1:t+\ell}^{\pi_t^*, \mu} = h_{t+1:t+\ell}^{\pi_t^*, \nu_t^*}$ , i.e.  $\nu_t^*$  is consistent with  $h_{t+1:t+\ell}^\circ$ , and hence  $\pi_t^*$  and  $\nu_t^*$  do not change from  $t+1, \dots, t+\ell$  (inner loop of Algorithm 1). Then

$$\begin{aligned}
& \begin{array}{ccccccc}
& \text{drop terms,} & & \text{same } h_{t+1:t+\ell}, & & \pi^\circ = \pi_t^* \text{ on } h_{t+1:t+\ell}, & \\
V_\mu^{\pi^\circ}(h_t) & \geq & V_{\mu,\ell}^{\pi^\circ}(h_t) & \stackrel{\circ}{=} & V_{\nu_t^*,\ell}^{\pi_t^*}(h_t) & \stackrel{\circ}{=} & V_{\nu_t^*,\ell}^{\pi_t^*}(h_t) \\
\geq & \uparrow & V_{\nu_t^*}^{\pi_t^*}(h_t) - \frac{\gamma^\ell}{1-\gamma} & = & \max_{\nu \in \mathcal{M}_t} \max_{\pi \in \Pi} V_\nu^\pi(h_t) - \varepsilon & \geq & \max_{\pi \in \Pi} V_\mu^\pi(h_t) - \varepsilon. \\
\text{bound extra terms} & & \text{def. of } (\pi_t^*, \nu_t^*) \text{ and } \varepsilon := \frac{\gamma^\ell}{1-\gamma} & & \mu \in \mathcal{M}_t & & 
\end{array}
\end{aligned}$$

Now let  $t_1, \dots, t_K$  be the times  $t$  at which the currently selected  $\nu_t^*$  gets inconsistent with  $h_t$ , i.e.  $\{t_1, \dots, t_K\} = \{t : \nu_t^* \notin \mathcal{M}_t\}$ . Therefore  $h_{t+1:t+\ell}^\circ \neq h_{t+1:t+\ell}^{\pi_t^*, \nu_t^*}$  (only) at times  $t \in \mathcal{T}_\times := \bigcup_{i=1}^K \{t_i - \ell, \dots, t_i - 1\}$ , which implies  $V_\mu^{\pi^\circ}(h_t) \geq \max_{\pi \in \Pi} V_\mu^\pi(h_t) - \varepsilon$  except possibly for  $t \in \mathcal{T}_\times$ . Finally

$$|\mathcal{T}_\times| = \ell \cdot K < \ell \cdot |\mathcal{M}| = \frac{\log \varepsilon(1-\gamma)}{\log \gamma} |\mathcal{M}| \leq |\mathcal{M}| \frac{\log \varepsilon(1-\gamma)}{\gamma - 1}$$

■

We refer to the algorithm above as the conservative agent since it sticks to its model for as long as it can. The corresponding liberal agent reevaluates its optimistic hypothesis at every time step and can switch between different optimistic policies at any time. Algorithm 1 is actually a special case of this as shown by Lemma 2. The liberal agent is really a class of algorithms and this larger class of algorithms consists of exactly the algorithms that are optimistic at every time step without further restrictions. The conservative agent is the subclass of algorithms that only switch hypothesis when the previous is contradicted. The results for the conservative agent can be extended to the liberal one, but we have to omit that here for space reasons.

### 3 Stochastic Environments

A stochastic hypothesis may never become completely inconsistent in the sense of assigning zero probability to the observed sequence while still assigning very different probabilities than the true environment. Therefore, we exclude based on a threshold for the probability assigned to the generated history. Unlike in the deterministic case, a hypothesis can cease to be the optimistic one without having been excluded. We, therefore, only consider an algorithm that reevaluates its optimistic hypothesis at every time step. Algorithm 2 specifies the procedure and Theorem 4 states that it is asymptotically optimal.

**Require:** Finite class of stochastic environments  $\mathcal{M}_1 \equiv \mathcal{M}$ , threshold  $z \in (0, 1)$

```

1:  $t = 1$ 
2: repeat
3:    $(\pi^*, \nu^*) = \arg \max_{\pi, \nu \in \mathcal{M}_t} V_\nu^\pi(h_{t-1})$ 
4:    $a_t = \pi^*(h_{t-1})$ 
5:   Perceive  $o_t r_t$  from environment  $\mu$ 
6:    $h_t \leftarrow h_{t-1} a_t o_t r_t$ 
7:    $t \leftarrow t + 1$ 
8:    $\mathcal{M}_t := \{\nu \in \mathcal{M}_{t-1} : \frac{\nu(h_t|a_{1:t})}{\max_{\tilde{\nu} \in \mathcal{M}} \tilde{\nu}(h_t|a_{1:t})} \geq z\}$ 
9: until the end of time

```

**Algorithm 2:** Optimistic Agent ( $\pi^\circ$ ) with Stochastic Finite Class

**Theorem 4** (Optimality, Finite Stochastic Class). *Define  $\pi^\circ$  by using Algorithm 2 with any threshold  $z \in (0, 1)$  and a finite class  $\mathcal{M}$  of stochastic environments containing the true environment  $\mu$ , then with probability  $1 - z|\mathcal{M} - 1|$  there exists, for every  $\varepsilon > 0$ , a number  $T < \infty$  such that*

$$V_\mu^{\pi^\circ}(h_t) > \max_{\pi} V_\mu^\pi(h_t) - \varepsilon \quad \forall t \geq T.$$

We borrow some techniques from [Hut09] that introduced a “merging of opinions” result that generalized the classical theorem by [BD62]. The classical result says that it is sufficient that the true measure (over infinite sequences) is absolutely continuous with respect to a chosen a priori distribution to guarantee that they will almost surely merge in the sense of total variation distance. The generalized version is given in Lemma 6. When we combine a policy  $\pi$  with an environment  $\nu$  by letting the actions be taken by the policy, we have defined a measure, denoted by  $\nu(\cdot|\pi)$ , on the space of infinite sequences from a finite alphabet. We denote such a sample sequence by  $\omega$  and the  $a$ :th to  $b$ :th elements of  $\omega$  by  $\omega_{a:b}$ . The  $\sigma$ -algebra is generated by the cylinder sets  $\Gamma_{y_{1:t}} := \{\omega | \omega_{1:t} = y_{1:t}\}$  and a measure is determined by its values on those sets. To simplify notation in the next lemmas we will write  $P(\cdot) = \nu(\cdot|\pi)$ , meaning that  $P(\omega_{1:t}) = \nu(h_t|a_{1:t})$  where  $\omega_j = o_j r_j$  and  $a_j = \pi(h_{j-1})$ . Furthermore,  $\nu(\cdot|h_t, \pi) = P(\cdot|h_t)$ .

**Definition 5** (Total Variation Distance). *The total variation distance between two measures (on infinite sequences  $\omega$  of elements from a finite alphabet)  $P$  and  $Q$  is defined to be*

$$d(P, Q) = \sup_A |P(A) - Q(A)|$$

where  $A$  is in the previously specified  $\sigma$ -algebra generated by the cylinder sets.

The results from [Hut09] are based on the fact that  $Z_t = \frac{Q(\omega_{1:t})}{P(\omega_{1:t})}$  is a martingale sequence if  $P$  is the true measure and therefore converges with  $P$  probability 1 [Doo53]. The crucial question is if the limit is strictly positive or not. The following lemma shows that with  $P$  probability 1 we are either in the case where

the limit is 0 or in the case where  $d(P(\cdot|\omega_{1:t}), Q(\cdot|\omega_{1:t})) \rightarrow 0$ . We say that the environments  $\nu_1$  and  $\nu_2$  merge under  $\pi$  if  $d(\nu_1(\cdot|\pi), \nu_2(\cdot|\pi)) \rightarrow 0$ .

**Lemma 6** (Generalized merging of opinions [Hut09]). *For any measures  $P$  and  $Q$  it holds that  $P(\Omega^\circ \cup \bar{\Omega}) = 1$  where*

$$\Omega^\circ := \{\omega : \frac{Q(\omega_{1:t})}{P(\omega_{1:t})} \rightarrow 0\} \quad \text{and} \quad \bar{\Omega} := \{\omega : d(P(\cdot|\omega_{1:t}), Q(\cdot|\omega_{1:t})) \rightarrow 0\}$$

**Lemma 7** (Value convergence for merging environments). *Given a policy  $\pi$  and environments  $\mu$  and  $\nu$  it follows that*

$$|V_\mu^\pi(h_t) - V_\nu^\pi(h_t)| \leq \frac{1}{1-\gamma} d(\mu(\cdot|h_t, \pi), \nu(\cdot|h_t, \pi)).$$

*Proof.* The lemma follows from the general inequality

$$|\mathbb{E}_P(f) - \mathbb{E}_Q(f)| \leq \sup |f| \cdot \sup_A |P(A) - Q(A)|$$

by inserting  $f := R(\omega_{t:\infty})$  and  $P = \mu(\cdot|h_t, \pi)$  and  $Q = \nu(\cdot|h_t, \pi)$ , and using  $0 \leq f \leq 1/(1-\gamma)$ . ■

The following lemma replaces the property for deterministic environments that either they are consistent indefinitely or the probability of the generated history becomes 0.

**Lemma 8** (Merging of environments). *Suppose we are given two environments  $\mu$  (the true one) and  $\nu$  and a policy  $\pi$  (defined e.g. by Algorithm 2). Let  $P(\cdot) = \mu(\cdot|\pi)$  and  $Q(\cdot) = \nu(\cdot|\pi)$ . Then with  $P$  probability 1 we have that*

$$\lim_{t \rightarrow \infty} \frac{Q(\omega_{1:t})}{P(\omega_{1:t})} = 0 \quad \text{or} \quad \lim_{t \rightarrow \infty} |V_\mu^\pi(h_t) - V_\nu^\pi(h_t)| = 0.$$

*Proof.* This follows from a combination of Lemma 6 and Lemma 7. ■

The next lemma tells us what happens after all the environments that will be removed have been removed but we state it as if this was time  $t = 0$  for notational simplicity.

**Lemma 9** (Optimism is nearly optimal). *Suppose that we have a (finite or infinite) class of (possibly) stochastic environments  $\mathcal{M}$  containing the true environment  $\mu$ . Also suppose that none of these environments are excluded at any time by Algorithm 2 ( $\pi^\circ$ ) during an infinite history  $h$  that has been generated by running  $\pi^\circ$  in  $\mu$ . Given  $\varepsilon > 0$  there is  $\tilde{\varepsilon} > 0$  such that*

$$V_\mu^{\pi^\circ}(\epsilon) \geq \max_{\pi} V_\mu^\pi(\epsilon) - \varepsilon$$

*if*

$$|V_{\nu_1}^{\pi^\circ}(h_t) - V_{\nu_2}^{\pi^\circ}(h_t)| < \tilde{\varepsilon} \quad \forall t, \forall \nu_1, \nu_2 \in \mathcal{M}.$$



*Proof. (Theorem 4)* Given a policy  $\pi$ , let  $P(\cdot) = \mu(\cdot|\pi)$  where  $\mu \in \mathcal{M}$  is the true environment and  $Q = \nu(\cdot|\pi)$  where  $\nu \in \mathcal{M}$ . Let the outcome sequence (the sequence  $(o_1 r_1), (o_2 r_2), \dots$ ) be denoted by  $\omega$ . It follows from Doob's Martingale inequality [Doo53] that for all  $z \in (0, 1)$

$$P(\sup_t \frac{Q(\omega_{1:t})}{P(\omega_{1:t})} \geq 1/z) \leq z, \quad \text{which implies} \quad P(\inf_t \frac{P(\omega_{1:t})}{Q(\omega_{1:t})} \leq z) \leq z.$$

This proves, using a union bound, that the probability of Algorithm 2 ever excluding the true environment is less than  $z|\mathcal{M} - 1|$ .

The limits  $\frac{\nu(h_t|\pi^\circ)}{\mu(h_t|\pi^\circ)}$  converge almost surely as argued before using the Martingale convergence theorem. Lemma 8 tells us that any given environment (with probability one) is eventually excluded or is permanently included and merge with the true one under  $\pi^\circ$ . The remaining environments does, according to (and in the sense of) Lemma 8, merge with the true environment. Lemma 7 tells us that the difference between value functions (for the same policy) of merging environments converges to zero. Since there are finitely many environments and the ones that remain indefinitely in  $\mathcal{M}_t$  merge with the true environment under  $\pi^\circ$ , there is for every  $\tilde{\varepsilon} > 0$  a  $T$  such that when following  $\pi^\circ$ , it holds for all  $t \geq T$  that

$$|V_{\nu_1}^{\pi^\circ}(h_t) - V_{\nu_2}^{\pi^\circ}(h_t)| < \tilde{\varepsilon} \quad \forall \nu_1, \nu_2 \in \mathcal{M}_t.$$

The proof is concluded by Lemma 9 in the case where the true environment remains indefinitely included which happens with probability  $z|\mathcal{M} - 1|$ . ■

## 4 Compact Classes

In this section we discuss infinite but compact classes of stochastic environments. First note that without further assumptions, asymptotic optimality can be impossible to achieve, even for countably infinite deterministic environments [LH11a]. Here we consider classes that are compact with respect to the total variation distance, or more precisely with respect to

$$\tilde{d}(\nu_1, \nu_2) = \max_{h, \pi} d(\nu_1(\cdot|h, \pi), \nu_2(\cdot|h, \pi))$$

where  $d$  is total variation distance from Section 3. An example is the class of Markov Decision Processes (or POMDPs) with a certain number of states. Algorithm 2 does need modification to achieve asymptotic optimality in the compact case. An alternative to modifying the algorithm is to be satisfied with reaching optimality within a pre-chosen  $\varepsilon > 0$ . This can be achieved by first choosing a finite covering of  $\mathcal{M}$  with balls of total variation radius less than  $\varepsilon(1 - \gamma)$  and use Algorithm 2 with the centers of these balls. To have an algorithm that for any  $\varepsilon > 0$  eventually achieves optimality within  $\varepsilon$  is a more demanding task. This is because we need to be able to say that the true environment will remain indefinitely in the considered class with a

given confidence. For this purpose we introduce a confidence radius inspired by MDP solving algorithms like MBIE [SL05] and UCRL [AO06]. We still use the notation  $\mathcal{M}_t$  as in Algorithm 2 and we define Algorithm 3 based on replacing it with a larger  $\tilde{\mathcal{M}}_t$ . If we do not do this the true environment is likely to be excluded.

**Definition 10** (Confidence radius). *We denote all environments within  $r_t^z$  from  $\mathcal{M}_t$  by*

$$\tilde{\mathcal{M}}_t := \{\nu \in \mathcal{M} \mid \exists \tilde{\nu} \in \mathcal{M}_t : \tilde{d}(\tilde{\nu}, \nu) \leq r_t^z\}.$$

*Given  $z > 0$  we say that  $r_t^z(h_t)$  is a  $p$ -confidence radius sequence if  $r_t^z(h_t) \rightarrow 0$  almost surely and if the true environment is in  $\tilde{\mathcal{M}}_t$  for all  $t$  with probability  $p$ .*

**Definition 11** (Algorithm 3). *Given a class of environments  $\mathcal{M}$  that is compact in the total variation distance we define Algorithm 3 as being Algorithm 2 with  $\mathcal{M}_t$  replaced by  $\tilde{\mathcal{M}}_t$*

**Definition 12** (Radon-Nikodym differentiable class). *Suppose that the class  $\mathcal{M}$  is such that if  $\mu \in \mathcal{M}$  is the true environment, then for any policy  $\pi$  it holds with probability one that for all  $\nu \in \mathcal{M}$ ,  $X_{t,\nu} := \frac{\nu(h_t|\pi)}{\mu(h_t|\pi)}$  converges as  $t \rightarrow \infty$  to some random variables  $X_\nu$ . We call such a class Radon-Nikodym (RN) differentiable. If the property holds with respect to a specific policy  $\pi$  we say that the class is RN-differentiable with respect to  $\pi$ .*

**Remark 13.** *Every countable class is RN-differentiable and so is the class of MDPs with a certain number of states. The MBIE [SL05] and UCRL [AO06] algorithms are based on the fact that one can define confidence radiuses for MDPs, though their bounds need separate intervals for each state-action pair depending on the number of visits. For an ergodic MDP all state-action pairs will almost surely be seen infinitely often and the max length of those intervals will tend to zero. Therefore, one can define a radius based on this maximum length or, alternatively, one can easily allow Algorithm 3 to run with such rectangular sets instead.*

**Theorem 14** (Optimality, Compact Stochastic Class). *Suppose we use Algorithm 3 with threshold  $z \in (0, 1)$ , a compact (in total variation) RN-differentiable class (with respect to  $\pi^\circ$  is enough)  $\mathcal{M}$  of stochastic environments and a  $p$ -confidence radius sequence  $r_t^z$  for  $\mathcal{M}$ . Denote the resulting policy by  $\pi^\circ$ . If the true environment  $\mu$  is in  $\mathcal{M}$ , then with probability  $p$  there is, for every  $\varepsilon > 0$ , a time  $T < \infty$  such that*

$$V_\mu^{\pi^\circ}(h_t) \geq \max_\pi V_\mu^\pi(h_t) - \varepsilon \quad \forall t \geq T.$$

**Lemma 15** (Uniform exclusion). *Let  $Q_\nu(\cdot) = \nu(\cdot|\pi^\circ)$  and  $P(\cdot) = \mu(\cdot|\pi^\circ)$  where  $\mu$  is the true environment and  $\pi^\circ$  the policy defined by Algorithm 3. For any outcome sequence  $\omega$ , let*

$$\mathcal{M}^0(\omega) := \{\nu \mid \frac{Q_\nu(\omega_{1:t})}{P(\omega_{1:t})} \rightarrow 0\}.$$

For any closed subset of  $\mathcal{M}^0(\omega)$  and for every  $z > 0$ , there is  $T < \infty$  such that for every  $\nu$  in this subset there is  $t \leq T$  such that  $\frac{Q_\nu(\omega_{1:t})}{P(\omega_{1:t})} < z$ .

*Proof.* Since  $\mathcal{M}$  is compact and the subset in question is closed it follows that it is also compact. Using the Arzelà-Ascoli Theorem [Rud76] we conclude that there is a subsequence  $t_k$  such that  $Z_k^\nu := \min\{1, \frac{Q_\nu(\omega_{1:t_k})}{P(\omega_{1:t_k})}\}$  converges uniformly to 0 on  $\mathcal{M}^0$  which means that there is  $t_k$  such that  $Z_k^\nu < z$  for all  $\nu \in \mathcal{M}^0$  and we can let  $t = T = t_k$ . ■

*Proof. (Theorem 14)* The strategy is to use that all environment that will be excluded and does not lie within a certain distance of some environment that merges with the true one, will be excluded after a certain finite time. Then we can say that the remaining environments' value functions differ at most by a certain amount and we can apply Lemma 9.

We can with probability one say that for each  $\nu \in \mathcal{M}$ , it will hold that  $Z_t = \frac{\nu(h_t|\pi^\circ)}{\mu(h_t|\pi^\circ)}$  converges and each environment will be in  $\mathcal{M}^0 = \{\nu \in \mathcal{M} \mid Z_t \rightarrow 0\}$  or  $\bar{\mathcal{M}} = \{\nu \mid d(\nu(\cdot|h_t, \pi^\circ), \mu(\cdot|h_t, \pi^\circ)) \rightarrow 0\}$ .  $\bar{\mathcal{M}}$  is compact (in the total variation distance topology) since it is a closed subset (again in the topology defined by  $\tilde{d}$ ) of the compact set  $\mathcal{M}$ .

For any  $\tilde{\varepsilon}_1 > 0$  we can do the following: For each  $\nu \in \mathcal{M}$ , consider a total variation ball of radius  $2\delta$  where  $\delta = (1 - \gamma)\tilde{\varepsilon}_1/4$ . Note that  $|V_\nu^{\pi^\circ}(h_t) - V_{\nu'}^{\pi^\circ}(h_t)| < \tilde{\varepsilon}_1/2$  for all  $t$  whenever  $\tilde{d}(\nu, \nu') < 2\delta$ . The collection of these balls induces an open cover of the compact set  $\mathcal{M}$  and it follows that there is a finite subcover. Consider the balls in this finite cover that intersect with  $\bar{\mathcal{M}}$ . Let  $\mathcal{A}$  be the union of these finitely many open balls. Let  $\mathcal{B} = \mathcal{M} \setminus \mathcal{A}$ .  $\mathcal{B}$  is then a closed subset of  $\mathcal{M}^0$ . We want to say that there is a finite time after which all environments in  $\mathcal{B}$  will have been excluded from  $\tilde{\mathcal{M}}_t$ . This happens if  $\tilde{\mathcal{B}}$ , defined as the union of the closed balls of radius  $r_t^z$  at every point in  $\mathcal{B}$ , has been excluded from  $\mathcal{M}_t$ . If  $t$  is large enough for  $r_t^z < \delta$ , then  $\mathcal{B}$  is also a closed subset of  $\mathcal{M}_0$ . Lemma 15 tells us that all of the environments in  $\tilde{\mathcal{B}}$  will have been excluded from  $\mathcal{M}_t$  after a finite amount of time  $T_1$  and, therefore, all the environments in  $\mathcal{B}$  will have been excluded from  $\tilde{\mathcal{M}}_t$ . Thus  $\tilde{\mathcal{M}}_t \subset \mathcal{A} \forall t \geq T_1$  and in particular the optimistic hypothesis  $\nu^*$  will be in  $\mathcal{A}$  when  $t \geq T_1$ . Let  $\nu^*(= \nu_t^*)$  be the optimistic hypothesis at time  $t \geq T_1$  and  $\pi^*(= \pi_t^*)$  the optimistic policy.

Each parameter in  $\mathcal{A}$  (and in particular  $\nu^*$ ) lies within  $\delta$  of a ball with center  $\nu$  which lies within  $\delta$  of a point  $\tilde{\nu} \in \bar{\mathcal{M}}$ . Hence  $\tilde{d}(\nu^*, \tilde{\nu}) < 2\delta$  and  $|V_{\nu^*}^{\pi^\circ}(h_t) - V_{\tilde{\nu}}^{\pi^\circ}(h_t)| < \tilde{\varepsilon}_1/2$ .

Due to the uniform merging of environments (under  $\pi^\circ$ ) on  $\bar{\mathcal{M}}$ , there is  $T_2 \geq T_1$  such that  $|V_{\nu_1}^{\pi^\circ}(h_t) - V_{\nu_2}^{\pi^\circ}(h_t)| < \tilde{\varepsilon}_1/2 \forall \nu_1, \nu_2 \in \bar{\mathcal{M}} \forall t \geq T_2$ . We conclude that  $|V_{\nu_1}^{\pi^\circ}(h_t) - V_{\nu_2}^{\pi^\circ}(h_t)| < \tilde{\varepsilon}_1 \forall \nu_1, \nu_2 \in \mathcal{A} \forall t \geq T_2$  and since  $\tilde{\mathcal{M}}_t \subset \mathcal{A}$

$$|V_{\nu_1}^{\pi^\circ}(h_t) - V_{\nu_2}^{\pi^\circ}(h_t)| < \tilde{\varepsilon}_1 \forall \nu_1, \nu_2 \in \tilde{\mathcal{M}}_t \forall t \geq T_2.$$

From Lemma 9 we know that if we picked  $\tilde{\varepsilon}_1$  small enough we know that for  $t \geq T_2$ ,  $V_{\nu^*}^{\pi^\circ}(h_t) \geq V_\nu^\pi(h_t) - \varepsilon/2$  for all  $\pi \in \Pi, \nu \in \mathcal{M}_t$ . Furthermore, by

picking  $\tilde{\varepsilon}_1$  sufficiently small we can, for  $t \geq T_2$ , ensure that there is  $\tilde{\nu} \in \tilde{\mathcal{M}}_t$  such that  $|V_{\tilde{\nu}}^{\pi^\circ}(h_t) - V_{\mu}^{\pi^\circ}(h_t)| < \varepsilon/2$ . Given that the true environment remains indefinitely in  $\tilde{M}_t$ , which happens with at least probability  $p$ , it follows that

$$V_{\mu}^{\pi^\circ}(h_t) \geq \max_{\pi} V_{\mu}^{\pi}(h_t) - \varepsilon \quad \forall t \geq T_2. \quad \blacksquare$$

## 5 Conclusions

We introduced optimistic agents for finite and compact classes of arbitrary environments and proved asymptotic optimality. In the deterministic case we also bound the number of time steps for which the value of following the algorithm is more than a certain amount lower than optimal. Future work includes investigating finite-error bounds for classes of stochastic environments.

**Acknowledgement.** This work was supported by ARC grant DP120100950. The authors are grateful for feedback from Tor Lattimore and Wen Shao.

## References

- [AO06] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Proceedings of NIPS'2006*, pages 49–56, 2006.
- [BD62] D. Blackwell and L. Dubins. Merging of Opinions with Increasing Information. *The Annals of Mathematical Statistics*, 33(3):882–886, 1962.
- [Doo53] J. Doob. *Stochastic processes*. Wiley, New York, NY, 1953.
- [EDKM05] E. Even-Dar, S. Kakade, and Y. Mansour. Reinforcement learning in pomdps without resets. In *Proceedings of IJCAI-05*, pages 690–695, 2005.
- [Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.
- [Hut09] M. Hutter. Discrete MDL predicts in total variation. In *Advances in Neural Information Processing Systems 22: (NIPS'2009)*, pages 817–825, 2009.
- [KS98] M. J. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. In *Proceedings of the 15<sup>th</sup> International Conference on Machine Learning (ICML'1998)*, pages 260–268, 1998.
- [LH11a] T. Lattimore and M. Hutter. Asymptotically optimal agents. In *Proc. of Algorithmic Learning Theory (ALT'2011)*, volume 6925 of *Lecture Notes in Computer Science*, pages 368–382. Springer, 2011.

- [LH11b] T. Lattimore and M. Hutter. Time consistent discounting. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, pages 383–397, Espoo, Finland, 2011. Springer, Berlin.
- [LH12] T. Lattimore and M. Hutter. PAC bounds for discounted MDPs. In *Proc. 23rd International Conf. on Algorithmic Learning Theory (ALT'12)*, volume 7568 of *LNAI*, Lyon, France, 2012. Springer, Berlin.
- [MMR11] O.-A. Maillard, R. Munos, and D. Ryabko. Selecting the state-representation in reinforcement learning. In *Advances in Neural Information Processing Systems 24 (NIPS'2011)*, pages 2627–2635, 2011.
- [Ors10] L. Orseau. Optimality issues of universal greedy agents with static priors. In *Proc. of Algorithmic Learning Theory, (ALT'2010)*, volume 6331 of *Lecture Notes in Computer Science*, pages 345–359. Springer, 2010.
- [RH08] D. Ryabko and M. Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theor. C.S.*, 405(3):274–284, 2008.
- [RN10] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, 3<sup>rd</sup> edition, 2010.
- [Rud76] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill, 1976.
- [SB98] R. Sutton and A. Barto. *Reinforcement Learning*. The MIT Press, 1998.
- [SL05] A. Strehl and M. Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of ICML 2005*, pages 856–863, 2005.